

COMPARISON OF ENSEMBLE-MOS METHODS USING GFS REFORECASTS

Daniel S. Wilks

*Department of Earth and Atmospheric Science
Cornell University, Ithaca, New York*

Thomas M. Hamill

*NOAA Earth System Research Laboratory
Boulder, Colorado*

Submitted to *Monthly Weather Review*

20 June 2006

Corresponding Author Address:

Dr. Daniel S. Wilks

Department of Earth and Atmospheric Science
Bradfield Hall, Cornell University, Ithaca, New York 14853
dsw5@cornell.edu, (607) 255-1750 (voice), 255-2106 (fax)

ABSTRACT

Three recently proposed and promising methods for post-processing ensemble forecasts based on their historical error characteristics, i.e., ensemble-MOS methods, are compared using a multi-decadal reforecast data set. Logistic regressions and non-homogeneous Gaussian regressions are generally preferred for daily temperature, and for medium-range (6–10 and 8–14 day) temperature and precipitation forecasts. However, the better sharpness of medium-range ensemble-dressing forecasts sometimes yields the best Brier scores even though their calibration is somewhat worse. Using the long (15- or 25-year) training samples that are available with these reforecasts improves the accuracy and skill of these probabilistic forecasts to levels that are approximately equivalent to gains of one day of lead time, relative to using short (1- or 2-year) training samples.

1. Introduction

Ensemble forecasts are now regularly produced by numerical weather prediction facilities worldwide (Toth and Kalnay 1993, 1997, Molteni et al. 1996, Houtekamer et al. 1996). The intent of ensemble forecasting is to provide a flow-dependent sample of the probability distribution of possible future atmospheric states. Ideally, the probability of any event could be skillfully estimated directly from the relative event frequency in the ensemble. Unfortunately, even when the ensemble has a small spread, so that the expected forecast uncertainty is small and the expected skill is large, the actual skill of such probabilistic forecasts may be much smaller than expected. Commonly, the forecasts are contaminated by systematic biases, and the ensemble spread is too small (e.g., Hamill and Colucci 1997, 1998, Buizza et al. 2005). These biases may be due to model errors due to insufficient resolution (Weisman et al. 1997, Mullen and Buizza 2002, Szunyogh and Toth 2002, Buizza et al. 2003) or sub-optimal parameterizations, sub-optimal methods for generating the initial conditions (Barkmeijer et al. 1998, 1999, Hamill et al. 2000, 2003, Wang and Bishop 2003, Sutton et al. 2006), the deterministic formulation of the forecast model (Palmer 2001, Wilks 2005), and other causes.

Consequently, many methods of calibrating the probabilistic forecasts from ensembles have been proposed. Most of these methods share a general approach of correcting the current forecast using past forecast errors, as has been done for deterministic forecasts in the Model Output Statistics, or “MOS” procedure (Glahn and Lowry 1972, Carter et al. 1989, Vislocky and Fritsch 1995, 1997, Krishnamurti et al. 1999). More recently, technique development has focused on probabilistic methods, including rank histogram techniques (Hamill and Colucci 1997, 1998, Eckel and Walters

1998), ensemble dressing (i.e., kernel density) approaches (Roulston and Smith 2003, Wang and Bishop 2005, Fortin et al. 2006), Bayesian model averaging (Raftery et al. 2005), non-homogeneous Gaussian regression (Gneiting et al. 2005), logistic regression (Hamill et al. 2004, Hamill and Whitaker 2006), analog techniques (Hamill et al. 2006, Hamill and Whitaker 2006), “forecast assimilation” (Stephenson et al. 2005), and several others.

If the systematic errors in the ensemble are consistent, then small training data sets may be adequate for correction of ensemble forecast errors. However, systematic errors may potentially vary from one synoptic situation to the next, or the small training data set may be inadequate for the forecast problem at hand. For example, if calibrating a 8–14 day average forecast, a month of training data will provide only barely four independent samples of training data. In such situations, a long training data set from a fixed numerical weather prediction model would be helpful. Recently, such an ensemble “reforecast” dataset was produced (Hamill et al. 2006) for a reduced-resolution, circa 1998 version of the National Centers for Environmental Prediction’s (NCEP’s) Global Forecast System (GFS). A 15-member ensemble reforecast has been produced out to 15 days lead for every day from 1979 to current. Skill improvements utilizing the long reforecast training data set have been demonstrated for 6-10 day and week-2 forecasts (Hamill et al. 2004, Whitaker et al. 2006), probabilistic quantitative precipitation forecasts (Hamill et al. 2006, Hamill and Whitaker 2006, Fortin et al. 2006), and hydrologic forecasts (Clark and Hay 2004, Gangopadhyay et al. 2004).

The reforecast data set described above offers an interesting opportunity to test a variety of the proposed methods for calibration of ensemble forecasts in a statistically

rigorous fashion. Recently, Wilks (2006a) compared a wide variety of calibration methods using the low-order Lorenz (1996) model (see also Lorenz and Emanuel 1998). The most promising approaches were logistic regression, non-homogeneous Gaussian regression (linear regression with non-constant prediction errors that depend on the ensemble spread), and ensemble dressing. Accordingly, in this article we shall compare these three techniques, by constructing probabilistic daily surface-temperature forecasts at lead times from 1 to 14 days, and average temperature and precipitation forecasts at 6–10 and 8–14 -day lead times. We will examine several questions. First, what is the relative performance of these techniques for producing post-processed probability forecasts from the reforecast data set? Second, does this relative performance change depending on the lead time, the lengths of the available training data, or other aspects of the forecast such as the forecast quantile? Third, how much absolute improvement is obtained by training the three ensemble-MOS methods with large vs. small samples?

The rest of the article will be organized as follows. Section 2 reviews the three ensemble-MOS methods to be compared, and Section 3 describes the reforecast and observational data to be employed. Section 4 outlines the experimental set-up, Section 5 presents cross-validated probabilistic verification results for the various forecasts, and Section 6 concludes.

2. Candidate ensemble-MOS methods

Wilks (2006a) evaluated a collection of ensemble-MOS methods that have been proposed in the literature using the low-order Lorenz (1996) model. The three most promising of these are described in this section, and are compared using the reforecast data set in Section 5.

2a. Logistic regression (LR)

The probability that a future observation, or verification V , will be less than or equal to a forecast quantile q can be specified using the 2-predictor logistic regression

$$\Pr(V \leq q) = \frac{\exp(b_0 + b_1 \bar{x}_{ens} + b_2 \bar{x}_{ens} s_{ens})}{1 + \exp(b_0 + b_1 \bar{x}_{ens} + b_2 \bar{x}_{ens} s_{ens})} . \quad (1)$$

Here, b_0 , b_1 , and b_2 are fitted constants, \bar{x}_{ens} refers to the ensemble-mean forecast, and s_{ens} refers to the ensemble spread, i.e., the standard deviation. This equation, referred to as LR(2) hereafter, produces an “S-shaped” prediction surface that is bounded by $0 < \Pr\{V \leq q\} < 1$ (e.g., Wilks 2006b). Wilks (2006a) used the ensemble spread as the second logistic regression predictor, but here the product of the ensemble mean and the ensemble spread has been used because it yielded slightly better results for the reforecast data. This form of logistic regression in Eq. (1) also has the appealing interpretation that it is equivalent to a 1-predictor logistic regression that uses the ensemble mean as the single predictor, but in which the regression parameter b_1 is itself a linear function of the ensemble standard deviation. Therefore, the steepness of the logistic function as it rises or falls with its characteristic “S” shape can increase with decreasing ensemble spread,

yielding sharper forecasts (more frequent use of extreme probabilities) when the ensemble spread is small.

Hamill et al. (2004), working with 6–10 and 8–14 -day forecasts of accumulated precipitation, found that the second predictor in Eq. (1) was not justified (i.e., did not improve forecast performance for independent data), and used the 1-predictor version of Eq. (1) in which $b_2 = 0$. This important special case of Eq. (1), with ensemble mean as the single predictor, will be referred to as LR(1) hereafter. For both LR(2) and LR(1), the regression functions are fit iteratively, using the method of maximum likelihood (e.g., Wilks 2006b).

2b. *Non-homogeneous Gaussian regression (NGR)*

Gneiting et al. (2005) proposed an extension to conventional linear regression, referred to here as non-homogeneous Gaussian regression (NGR). The approach is to construct a conventional regression equation using ensemble mean as the single predictor, but to allow the variance characterizing the prediction uncertainty to vary as a linear function of the ensemble variance. That is, the variances of the regression errors are non-homogeneous (not the same for all values of the predictor), as is conventionally assumed in linear regression. Assuming also that the forecast uncertainty is adequately described by a Gaussian distribution leads to forecast probability estimation using

$$\Pr(V \leq q) = \Phi \left[\frac{q - (a + b\bar{x}_{ens})}{(c + d s_{ens}^2)^{1/2}} \right] . \quad (2)$$

Here, a and b are the linear regression intercept and slope, and c and d are parameters relating the prediction variance to the ensemble variance. The symbol Φ indicates the

cumulative distribution function of the standard Gaussian distribution, and the quantity in the square brackets is a standardized variable, or “z-score” (i.e., a forecast quantile q minus its regression mean, divided by the prediction standard deviation), so that Eq. (2) yields forecast probability distributions that are explicitly Gaussian. Following Gneiting et al. (2005), the four parameters in Eq. (2) are fit iteratively, in order to minimize the continuous ranked probability score (e.g., Wilks 2006b) for the training data.

Equation (2) reduces to conventional ordinary least squares (OLS) predictions when the denominator is equal to the overall, constant prediction standard deviation ($c \approx \text{MSE}$ and $d = 0$). This important special case is also considered in Section 5.

2c. *Gaussian ensemble dressing (GED)*

The method of ensemble dressing (Roulston and Smith 2003, Wang and Bishop 2005) constructs an overall forecast probability distribution by centering probability distributions at each of the (de-biased) ensemble members, and then averaging these n_{ens} probability distributions. Ensemble dressing is thus a kernel density smoothing (e.g., Wilks 2006b) approach. When the smoothing kernels are Gaussian distributions, then the method is known as Gaussian ensemble dressing (GED), and the resulting forecasts for quantiles q are calculated as

$$\Pr(V \leq q) = \frac{1}{n_{\text{ens}}} \sum_{i=1}^{n_{\text{ens}}} \Phi \left[\frac{q - \tilde{x}_i}{\sigma_D} \right] , \quad (3)$$

where the tilde denotes that any overall bias in the training data has been removed from each ensemble member x_i . That is, a single correction, equal to the average difference between the ensemble means and their corresponding verifications in the training data, is

applied equally to all ensemble members, so that the ensemble dispersion is not affected. Note that even though the dressing kernels are specified as Gaussian, the overall forecast distribution is in general not Gaussian, and indeed can take on any shape that might be indicated by the distribution of the underlying ensemble members.

The key parameter in Eq. (3) is the standard deviation of the Gaussian dressing kernel, σ_D . Roulston and Smith (2003) propose fitting this parameter according to the forecast errors of the “best” member in each ensemble, although in real forecast situations, definition of this best member can be problematic. Here we use the Gaussian dressing variance proposed by Wang and Bishop (2005),

$$\sigma_D^2 = \sigma_{\bar{x}-y}^2 - \left(1 + \frac{1}{n_{ens}}\right) \bar{\sigma}_{ens}^2, \quad (4)$$

which is calculated as the difference between the error variance for the ensemble-mean forecasts and the (slightly inflated) average of the ensemble variances, over the training data. Equation (4) can sometimes fail (i.e., yield negative dressing variances) if the forecast ensembles in the training data are sufficiently over-dispersed, on average. In this study, this difficulty occurred only rarely in the training data; and in these cases Eq. (4) was formally set to zero, implying that all probability is assigned to the n_{ens} de-biased points, which is equivalent to estimating forecast probability using (de-biased) ensemble relative frequency (the “democratic voting” method).

3. Forecast and verification data

Ensemble forecasts for twice-daily temperature and precipitation were taken from the GFS reforecast data set (Hamill et al. 2006), for the period January 1979 through February 2005. Verification data are observed maximum temperature, minimum temperature and 24-hour accumulated precipitation at 19 midnight-observing, first-order U.S. National Weather Service stations: Atlanta (ATL), Bismarck (BIS), Boston, (BOS), Buffalo (BUF), Washington, D.C. (DCA), Denver (DEN), Dallas (DFW), Detroit (DTW), Great Falls (GTF), Los Angeles (LAX), Miami (MIA), Minneapolis (MSP), New Orleans (MSY), Omaha (OMA), Phoenix (PHX), Seattle (SEA), San Francisco (SFO), Salt Lake City (SLC), and St. Louis (STL). The reforecast data are available on a $2.5^\circ \times 2.5^\circ$ grid, and grid points nearest each of the 19 first-order stations were selected.

Two types of probabilistic forecasts are considered: daily maximum and minimum temperature forecasts, and medium-range (6–10 and 8–14 -day average) temperature and precipitation forecasts. Ensemble forecasts for near-surface (2 m a.g.l.) temperature, and accumulated precipitation, are available from the reforecast data set for 0000 and 1200 UTC, only. The calibration of maximum and minimum temperatures is an especially challenging application of this data set, as the daily maximum and minimum temperatures typically occur at times different than 0000 and 1200 UTC.

Probabilistic forecasts for daily maximum and minimum temperatures were made for lead times of 1, 2, 3, 5, 7, 10, and 14 days; and pertain to the following seven quantiles: $q_{.05}$ (5th percentile), $q_{.10}$ (lower decile), $q_{.33}$ (lower tercile), $q_{.50}$ (median), $q_{.67}$ (upper tercile), $q_{.90}$ (upper decile), and $q_{.95}$ (95th percentile). These quantiles were defined locally, both in time and individually for each verifying station, to avoid artificial

skill deriving from correct “forecasting” of variations in these climatological values (Hamill and Juras, 2006).

For the daily temperature forecasts, the cooler of these two twice-daily forecast temperatures (usually the 1200 UTC value) during each midnight-to-midnight observing period was assigned as the predictor for minimum temperature. The warmer of the two was assigned as the maximum temperature predictor. These assignments were made separately for each of the $n_{\text{ens}} = 15$ ensemble members.

For the 6–10 and 8–14 -day temperature forecasts, the twice-daily temperature forecasts were averaged, and the twice-daily precipitation forecasts were summed, separately for each ensemble member, over the respective lead times. For these medium-range forecast probability forecasts, the two terciles, $q_{.33}$ and $q_{.67}$ only, are considered.

4. Experimental set-up

Forecast equations were fit using 1, 2, 5, 15, and 25 years of training data, and evaluated using cross validation. For each forecast method described in Section 2, new forecast equations were fit for each day of the 26-year data period, using training-data windows of ± 15 , ± 30 , and ± 45 days around the corresponding date in each of the training years. To the extent possible, training years were chosen as those immediately preceding the year omitted for cross validation, and to the extent that this was not possible the nearest subsequent years were used. For example, using 1 year of training data and a ± 15 -day window, initial dates of training data for forecasts initialized on 1 March 1980, were 14 February – 16 March 1979. For forecasts initialized on 1 March 1980 but using

2 years of training data, data from these same initial dates in both 1979 and 1981 were used for training.

In addition, for the daily temperature forecasts, a “0-year” training strategy was tested, which is meant to simulate operational approaches to continuously updating MOS equations using only the most recent data (e.g., Wilson and Valée 2002, 2003). Here training data are taken only from the most recent 45 days available for each lead time, and so include only initial dates beginning 45 days (for the 1-day lead time) to 58 days (for the 14-day lead time) earlier.

5. Results

5a. Daily temperature forecasts

Figure 1 shows the cross-validated ranked probability scores (RPS) for the daily temperature forecasts, using the seven forecast temperature quantiles listed in Section 3. For clarity, only results for the 1-year and 25-year training periods are contrasted, and in each case only results for the training window yielding the best scores are shown. Qualitatively, results for minimum temperature forecasts (Fig. 1a) and maximum temperature forecasts (Fig. 1b) are similar, so both here and subsequently the discussion will focus on the minimum temperature forecasts.

Results for the longer training period (solid lines) are clearly superior (lower RPS) to those for the short training period (dashed lines). Best results for the 1-year training period are obtained for the longest (± 45 , or 91 days) training window, whereas when many years of training data are available the best results are obtained with the climatologically more focused short (± 15 , or 31 days) training window. Similarly, the

more elaborate LR(2) and NGR models in Eqs. (1) and (2), respectively, are not supported by (i.e., are overfit when using) the short 1-year training period, so that the simpler LR(1) and OLS special cases are chosen as best. In contrast, the longer training period provides sufficient data for the more elaborate LR(2) and NGR models to be usefully applied.

For the 25-year training period, logistic regressions provide a small but consistent improvement over the linear regressions (NGR), in terms of overall RPS. Both yield better RPS than the climatological probabilities (RPS_{clim}), over the entire 2-week forecast period. In contrast, for the short training period, the linear (OLS) regressions yield better RPS than the logistic [LR(1)] regressions. In no case do the GED forecasts yield the best overall results here, and the improvement in RPS for the GED forecasts between the 1- and 25-year training samples is small. Overall, and in particular for the linear and logistic regressions, differences in training lengths appear to be more important than differences in the forecast methods.

Figure 2 provides a somewhat different perspective on the overall RPS values for the daily minimum temperatures. Here cross-validated RPS for the best combination of forecast method (indicated by the plotting symbols) and training window (indicated parenthetically) is shown as a function of the training length. Here “0 years” training length denotes fitting the forecast equations using the preceding 45 days, only, as the training period. Using the best combination of forecast method and training window, the forecasts improve over the climatological RPS, except for day-14 forecasts made using the shortest training periods. Again, the linear regressions perform best for the shorter training periods, whereas the logistic regressions are preferred for the longer training

lengths, although as indicated in Figure 1 these differences among the forecast methods are often slight. There appears to be very little improvement in RPS when training data is increased from 15 to 25 years, and for both of these training lengths, the shortest (31-day) training window yields the best results. Using 15 or 25 years training data gains approximately one day of lead time in terms of RPS, relative to the shorter training lengths.

Fewer years in the training sample can be only partly compensated through use of wider training windows. For the most part, best results for 2- and 5-year training periods are obtained with 61-day windows, and best results for the 1-year training period are usually obtained using the 91-day training window. Training on the previous 45 days only (“0 years”) yields results that are quite similar to the 1-year training period (although with wider training windows), in terms of this overall accuracy measure.

Ranked probability scores provide a convenient single-number summary of forecast performance, but also combine and obscure some important details. Figure 3 shows a partial disaggregation of the RPS for the minimum temperature forecasts, in terms of Brier scores for $\Pr\{V \leq q_{0.05}\}$ (Fig. 3a) and $\Pr\{V \leq q_{0.33}\}$ (Fig. 3b). In general, results for forecasts of the lower tercile (Fig. 3b), which are representative of forecasts for other mid-distribution quantiles, are similar to the overall RPS values for daily minimum temperature forecasts shown in Fig. 2. In particular, the two-predictor LR(2) logistic regressions are generally best for the longer training samples, linear regressions (although usually NGR rather than OLS) are preferred for the shorter training samples, results for 15- and 25-year training periods are similar, and use of these longer training periods improves over results for the shorter training periods sufficiently to gain approximately

one day of lead time. Most of these observations hold also for Brier scores for the 5th percentiles (Fig. 3a), which are representative of those for other extreme quantiles, except for results regarding the best forecast method. Here the NGR linear regression method is justified in most cases, regardless of the training sample size.

A yet more detailed comparison of the forecast methods can be obtained from reliability diagrams for probability forecasts of particular quantiles. Figure 4 shows representative examples, for 2-day ahead forecasts of the lower terciles of minimum temperature, for the (October – March) cool season, using 1 year (Fig. 4a) and 15 years (Fig. 4b) of training data. For 1 year of training data the best-calibrated forecasts are clearly the linear regressions. Here the OLS forecasts are slightly more reliable (using the Murphy, 1973, decomposition of the Brier score, as shown in the inset) than the NGR forecasts. Reliability of the LR(1), LR(2) and GED forecasts are clearly inferior, and in particular exhibit over-forecasting for the higher probabilities. Interestingly, the GED forecasts yield the best Brier score in this case, as a result of their use of the higher probabilities more frequently, yielding a higher resolution (RES) component of the Brier score decomposition, even though they are the least well calibrated. For the longer, 15-year training sample (Fig. 4b), the best forecasts overall are provided by the LR(2) method. All of the forecast methods show improvement over results from Fig. 4a although, consistent with Fig. 1, the GED forecasts improve least.

Figure 5 shows reliability diagrams for day-2 cool-season forecasts of minimum temperature 5th percentiles. Overall, the relative results are similar to those in Fig. 4, although the calibrations are notably poorer, except for the linear regressions using 15 years of training data. The linear regressions are preferred for both training lengths, the

15-year training sample in Fig. 5b is not sufficient for the logistic regressions to produce fully calibrated forecasts, and of course the higher probabilities are used much less frequently for this extreme low quantile.

5b. Medium-range temperature and precipitation forecasts

Table 1 summarizes the broad features of the skill of the medium-range tercile probability forecasts, again made using the methods described in Section 2. These cross-validated results are for the ranked probability skill score, calculated relative to the climatological probabilities of $\Pr\{V \leq q_{1/3}\} = 1/3$ and $\Pr\{V \leq q_{2/3}\} = 2/3$, and contrasting the 1-year versus the 25-year training periods. Again the best training windows in each case have been chosen, which are 91 days for the 1-year training and for precipitation forecasts with 25 years of training data, and 31 days for temperature forecasts with 25 years of training data. Best results in each case are indicated in boldface.

Clearly the results are quite poor with only one year of training data, especially for the precipitation forecasts, for which all skills are negative. In several cases the GED forecasts yield the least bad results in this limited-data situation. With ample training data, the NGR forecasts are best for temperature, although the OLS and LR(1) forecasts are nearly as good. In contrast the linear regression forecasts are quite poor for precipitation, which is not surprising given that their forecast distributions are explicitly Gaussian. Here the best forecasts according to this overall measure are those made with the single-predictor LR(1) method. Consistent with the results obtained by Hamill et al. (2004), the 2-predictor LR(2) logistic regressions do not provide an improvement for these medium-range forecasts.

Figure 6 provides more detail on the performance of the medium-range forecasts. Here Brier scores are shown as functions of the length of the training data for the best forecast methods in each instance. As was the case for daily temperature forecasts, there is little improvement as the training length increases from 15 to 25 years (see also Hamill et al. 2004). The best temperature forecasts generally result from linear regressions, and from the NGR method in particular for the longer training lengths. The LR(1) method yields best Brier scores for forecasts of the upper tercile of precipitation, but the GED forecasts exhibit slightly better Brier scores for the lower tercile of precipitation, and for the lower-tercile 8–14 day temperature forecasts.

The reliability diagrams for the 6–10 day precipitation forecasts for October–March in Fig. 7 illustrate the reason for the good Brier scores exhibited by the GED forecasts in Fig. 6. Here the inset tables indicate that the GED forecasts yielded the best Brier scores overall (note that Fig. 6 shows full-year, not cool-season, results) for both terciles, yet are notably less well calibrated than forecasts from either of the logistic regression methods. This apparent discrepancy is explained by comparing the inset bar charts (note logarithmic vertical scales) showing frequencies of use of the forecast probabilities. For the LR(1) forecasts these are concentrated near the climatological values of $1/3$ (Fig. 7a) and $2/3$ (Fig. 7b); whereas the GED forecasts are much sharper, as these distributions of forecast useages are much more nearly uniform. Thus, even though the calibration of the GED forecasts is not as good, the Brier score credits them for their increased sharpness. It is possible that in this case some forecast users would find the GED forecasts more valuable, and that others might find the LR(1) forecasts more valuable (e.g., Ehrendorfer and Murphy 1988). The inset tables indicate that Brier scores

for the LR(1) forecasts are nearly as good as those for the GED forecasts. In contrast, neither of the Gaussian linear regression forecasts exhibit positive skill relative to the climatological probabilities ($BS_{\text{clim}} = 0.2222$).

5. Summary and Conclusions

This study has used the reforecast data set (Hamill et al. 2006) to compare three promising methods for ensemble-MOS forecasting identified in Wilks (2006a). The three methods are logistic regression (e.g., Hamill et al. 2004, Wilks 2006b), non-homogeneous Gaussian regression (Gneiting et al. 2005), and Gaussian ensemble dressing (Roulston and Smith 2003, Wang and Bishop 2005). The methods were tested for probabilistic forecasts of daily temperature at lead times of 1 to 14 days, and for 6–10 and 8–14 -day averages of both temperature and precipitation.

Reinforcing the results of Hamill et al. (2004), it was found that the longer (15- and 25-year) training samples available in the reforecast data set provide substantial forecast skill increases over short (1-, 2-, or 5-year) training periods, including use of the preceding 45 days only (“0 years”) as a training period. In particular, use of the longer training periods gains approximately one day of lead time, in terms of the accuracy and skill metrics employed, relative to the shorter training samples.

There appears to be no single best forecast method for all applications, among the three tested. When long training samples were available, the LR(2) (2-predictor logistic regression) method yielded the best RPS overall for daily temperature forecasts, and the best Brier scores for central forecast quantiles. However, the NGR forecasts exhibited slightly greater accuracy for probability forecasts of the more extreme daily temperature

quantiles. For 6–10 and 8–14 -day temperatures, the NGR forecasts were generally best. For the longer-range precipitation forecasts, the single-predictor logistic regressions were often best, as was found by Hamill et al. (2004), although in some instances the much better sharpness of the GED forecasts compensated for their poorer calibration to yield better Brier scores. Overall, differences in training lengths usually produced larger skill differences than did different forecast methods.

The combined results of Wilks (2006a) and the present paper cannot be regarded as the last word on ensemble-MOS methods. For example, Fortin et al. (2006) have recently proposed an ensemble-dressing method in which different members of the ranked ensemble may use different dressing kernels. Another possible extension of ensemble dressing could be to dress regression-corrected ensemble members (i.e., defining \tilde{x}_i in Eq. (3) as the result of a linear regression), which would yield a method similar to NGR, but with nonparametric forecast distributions. Similarly, this research does not make clear which ensemble-MOS method will perform best in calibrating other forecast variables, such as wind speed or direction, cloud cover, precipitation type, etc. Investigating such questions should be part of an overall program of development of calibrated probabilistic prediction system.

The judgment regarding whether operational use of reforecasts is worthwhile is ultimately a subjective, managerial one. However, the improved skill from calibration using large data sets is equivalent to the skill increases afforded by perhaps 5-10 years of numerical modeling system development and model resolution increases. While computationally expensive, the reforecasts may offer a comparatively inexpensive way of achieving increases in forecast skill. Hamill et al. (2006, conclusion section) discuss

some possible ways that reforecasts can be implemented into operations without unduly affecting the model development and production.

Acknowledgments

The second author's participation was partially supported by NSF grants ATM-0130154 and ATM-0205612. Jeff Whitaker and Xuguang Wang are thanked for their consultations during the production of this manuscript.

References

- Barkmeijer, J., M. van Gijzen, and F. Bouttier, 1998: Singular vectors and estimates of the analysis error covariance metric. *Quart. J. Royal Meteor. Soc.*, **124**, 1695-1713.
- , R. Buizza, and T.N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quart. J. Royal Meteor. Soc.*, **125**, 2333-2351.
- Buizza, R., D. S. Richardson, and T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble system and comparisons with poor-man's ensembles. *Quart. J. Roy. Meteor. Soc.*, **129**, 1269-1288.
- , P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076-1097.
- Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, **12**, 581-594.
- Clark, M. P., and L. E. Hay, 2004: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *J. Hydrometeor.*, **5**, 15-32.
- Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132-1147.

- Ehrendorfer, M., and A.H. Murphy, 1988: Comparative evaluation of weather forecasting systems: sufficiency, quality and accuracy. *Mon. Wea. Rev.* **116**, 1757-1770.
- Fortin, V. A.-C. Favre, and M. Saïd, 2006: Probabilistic forecasting from ensemble prediction systems: improving upon the best-member method by using a different weight and dressing kernel for each member. *Quart. J. Royal Meteor. Soc.*, in press. Available from <http://tinyurl.com/hvvrX>.
- Gangopadhyay, S., M.P. Clark, B. Rajagopalan, K. Werner, and D. Brandon, 2004: Effects of spatial and temporal aggregation on the accuracy of statistically downscaled precipitation estimates in the Upper Colorado River basin. *J. Hydrometeor.*, **5**, 1192-1206.
- Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.
- Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman. 2005: Calibrated probabilistic forecasting using ensemble Model Output Statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118.
- Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312-1327.
- , and -----, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711-724.
- , and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc.*, conditionally accepted. Available from www.cdc.noaa.gov/people/tom.hamill/cv.html.

- C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular vector, and perturbed observation ensembles. *Mon. Wea. Rev.*, **128**, 1835-1851.
- , ----- and J. S. Whitaker, 2002: Ensemble forecasts and the properties of flow-dependent analysis-error covariance singular vectors. *Mon. Wea. Rev.*, **131**, 1741-1758.
- , and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.*, in press. Available from <http://tinyurl.com/zoxoa> .
- , -----, and S. L. Mullen, 2006: Reforecasts, an important new data set for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33-46.
- , -----, and X. Wei, 2004: Ensemble re-forecasting: improving medium range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434-1447.
- Houtekamer, P. L., L. Lefaiivre, and J. Derome, 1996: The RPN ensemble prediction system. *Proc. ECMWF Seminar on Predictability, Vol II*, Reading, United Kingdom, 121-146. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom].
- Krishnamurti, T. N., and coauthors, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548-1550.
- Lorenz, E. N., 1996: Predictability: a problem partly solved. *Proc. ECMWF Seminar on Predictability, Vol 1*, 1-18. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

- , and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: simulations with a small model. *J. Atmos. Sci.*, **55**, 399-414.
- Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Royal Meteor. Soc.*, **122**, 73-119.
- Mullen, S. L., and R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173-191.
- Murphy, A.H., 1973: A new vector partition of the probability score. *J. Appl. Meteorol.* **12**, 595-600.
- Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: a proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quart. J. Roy. Meteor. Soc.*, **127**, 279-304.
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155-1174.
- Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16-30.
- Stephenson, D. B., C. A. S. Coelho, M. Balmaseda, and F. J. Doblas-Reyes, 2005: Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus A*, **57**, 253-264.

- Sutton, C. J., T. M. Hamill, and T. T. Warner, 2006: Will perturbing soil moisture improve warm-season ensemble forecasts? A proof of concept. *Mon. Wea. Rev.*, in press. Available at esrl.noaa.gov/psd/people/tom.hamill/land_sfc_perts_v2.pdf
- Szunyogh, I., and Toth, Z. 2002: The effect of increased horizontal resolution on the NCEP global ensemble mean forecasts. *Mon. Wea. Rev.*, **130**, 1125-1143.
- Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.* **74**, 2317-2330.
- , and -----, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.
- Vislocky, R. L., J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.* **76**, 1157-1164.
- , and -----, 1997: Performance of an advanced MOS system in the 1996-97 national collegiate weather forecasting contest. *Bull. Amer. Meteor. Soc.*, **78**, 2851-2857.
- Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140-1158.
- , and -----, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Royal Meteor. Soc.*, **131**, 965-986.
- Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527-548.
- Whitaker, J. S., X. Wei., and F. Vitart, 2006: Improving week-two forecasts with multi-model re-forecast ensembles. *Mon. Wea. Rev.*, in press. Available from Jeffrey.s.whitaker@noaa.gov.

Wilks, D. S., 2005: Effects of stochastic parameterization in the Lorenz '96 system.

Quart. J. Royal Meteor. Soc., **131**, 389-407.

-----, 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting.

Meteor. Apps., in press. Available from dsw5@cornell.edu.

-----, 2006b: *Statistical Methods in the Atmospheric Sciences*, 2nd Ed., Academic Press, 627 pp.

Wilson, L.J., and M. Valée, 2002: The Canadian updateable model output statistics

(UMOS) system: design and development tests. *Weather Forecasting*, **17**, 206-222.

-----, and -----, 2003: The Canadian updateable model output statistics

(UMOS) system: validation against perfect prog. *Wea. Forecasting*, **18**, 288-302.

Table 1. Percent RPS skill, relative to the climatological probabilities, for tercile forecasts of temperature and precipitation, at lead times of 6–10 and 8–14 days. Best skills in each case are indicated in bold.

| | 1 year | | | | 25 years | | | |
|--------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | Training | | Training | | Training | | Training | |
| | <u>T 6–10</u> | <u>T 8–14</u> | <u>P 6–10</u> | <u>P 8–14</u> | <u>T 6–10</u> | <u>T 8–14</u> | <u>P 6–10</u> | <u>P 8–14</u> |
| LR(1) | 3.2 | –8.9 | –5.2 | –9.4 | 14.5 | 7.4 | 2.9 | 1.3 |
| LR(2) | –16.2 | –32.2 | –19.1 | –26.0 | 10.3 | 2.1 | 1.8 | –0.3 |
| NGR | 7.9 | –1.4 | –15.2 | –14.9 | 14.7 | 7.7 | –10.1 | –8.3 |
| OLS | 8.1 | –1.1 | –17.4 | –17.5 | 14.5 | 7.6 | –13.2 | –10.9 |
| GED | 3.5 | 1.1 | –1.6 | –3.7 | 6.7 | 5.9 | 1.5 | –0.6 |

Figure Captions

Figure 1. Cross-validated ranked probability scores for daily (a) minimum, and (b) maximum temperatures, as functions of lead time, for 1- and 25 years of training data. Scores for the training window yielding best results in each case are shown.

Figure 2. Cross-validated RPS as a function of training length, for best combinations of forecast methods and training windows. “0 years” training length indicates use of the preceding 45 days, only, for the training period.

Figure 3. Brier scores for daily temperature forecasts of (a) 5th percentiles, and (b) lower terciles, as functions of years of training data. Best combinations of forecast method and training window are shown. “0 years” training length indicates training on the preceding 45 days only.

Figure 4. Reliability diagrams for day-2 October–March forecasts of the lower tercile of minimum temperature distributions, using (a) 1 year, and (b) 15 years of training data. Insets show frequencies of use of the forecasts for the best calibrated method in each case, and terms in the Murphy (1973) decomposition of the Brier score for each forecast method.

Figure 5. As Figure 4, for forecasts of the 5th percentile of cool-season minimum temperature.

Figure 6. Brier scores for medium-range forecasts of outcomes below (a) the lower tercile, and (b) the upper tercile, of the climatological distributions, as functions of the length of the training data. Best forecast methods and training windows are shown in each case.

Figure 7. As Figure 4, for 6–10 day forecasts that precipitation is at or below the (a) lower terciles, and (b) upper terciles, using 15 years of training data.

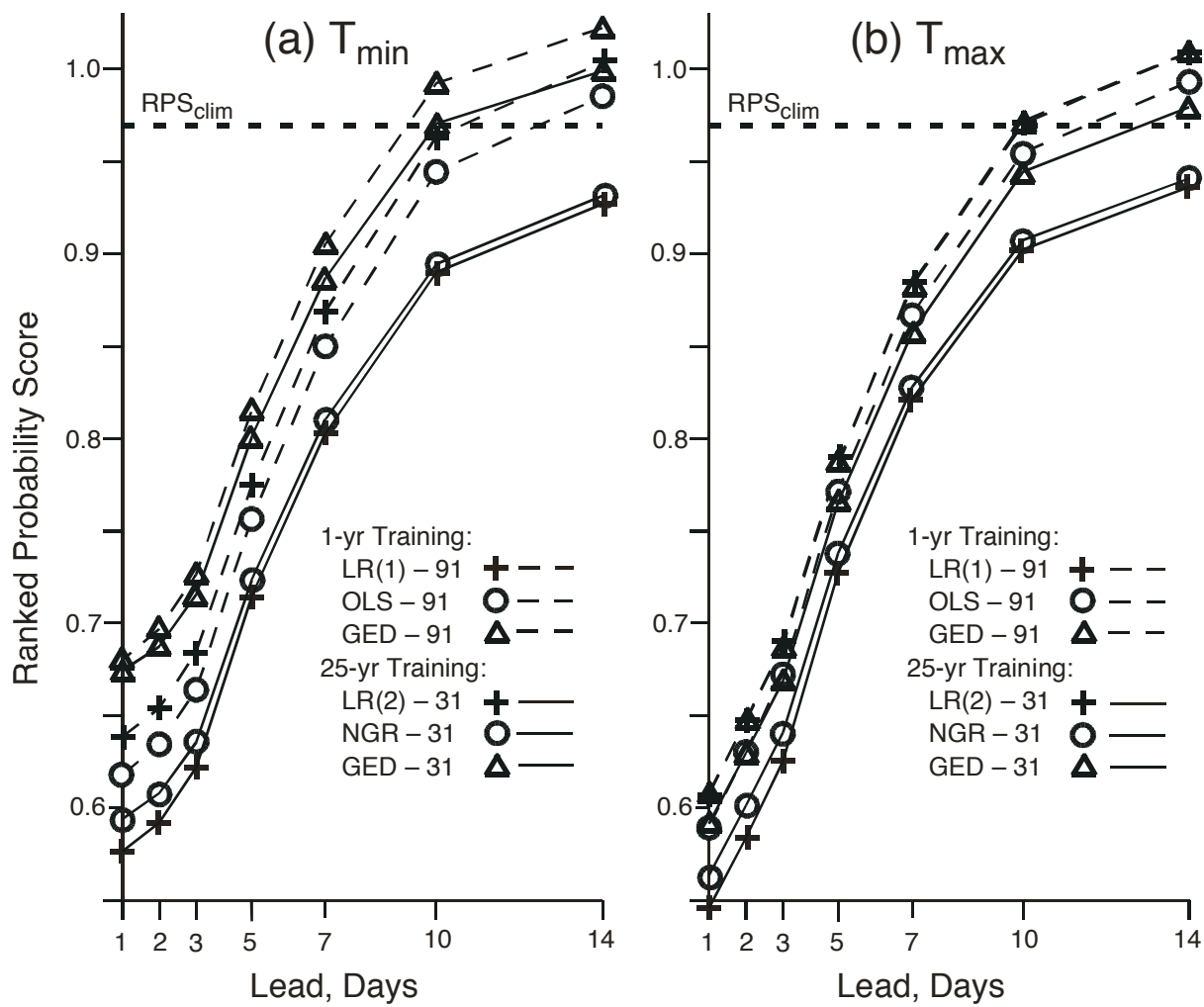


Figure 1. Cross-validated ranked probability scores for daily (a) minimum, and (b) maximum temperatures, as functions of lead time, for 1- and 25 years of training data. Scores for the training window yielding best results in each case are shown.

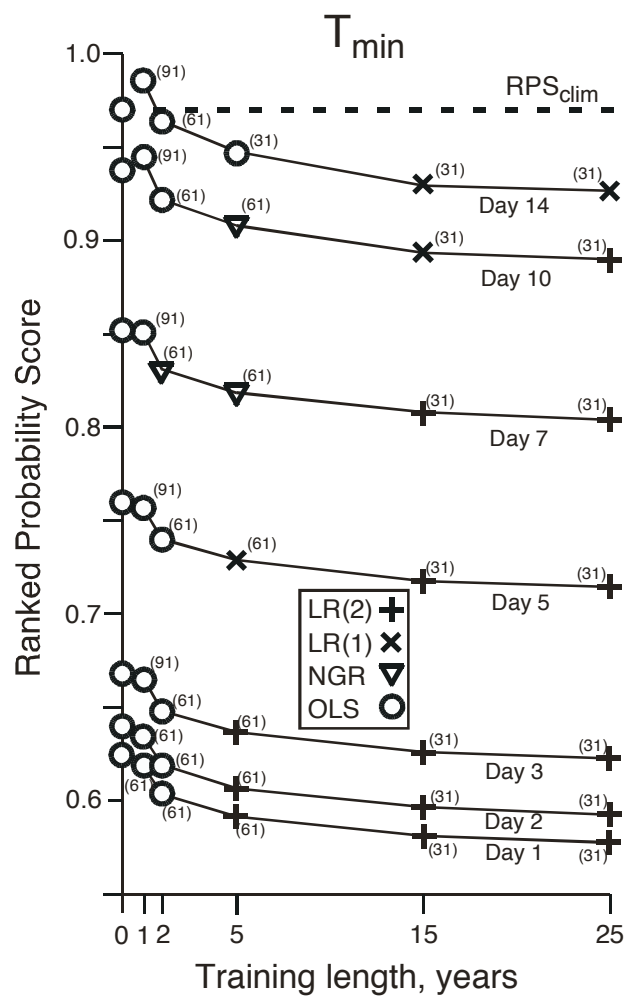


Figure 2. Cross-validated RPS as a function of training length, for best combinations of forecast methods and training windows. “0 years” training length indicates use of the preceding 45 days, only, for the training period.

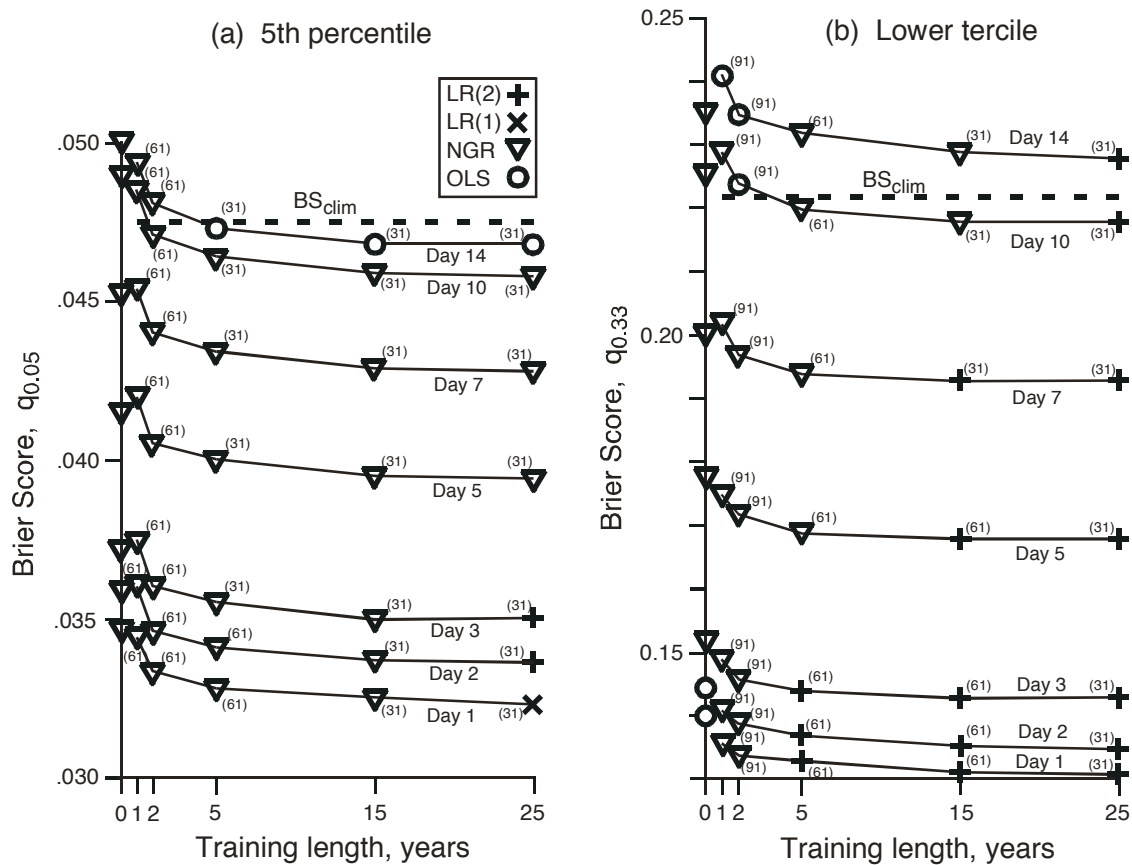


Figure 3. Brier scores for daily temperature forecasts of (a) 5th percentiles, and (b) lower terciles, as functions of years of training data. Best combinations of forecast method and training window are shown. “0 years” training length indicates training on the preceding 45 days only.

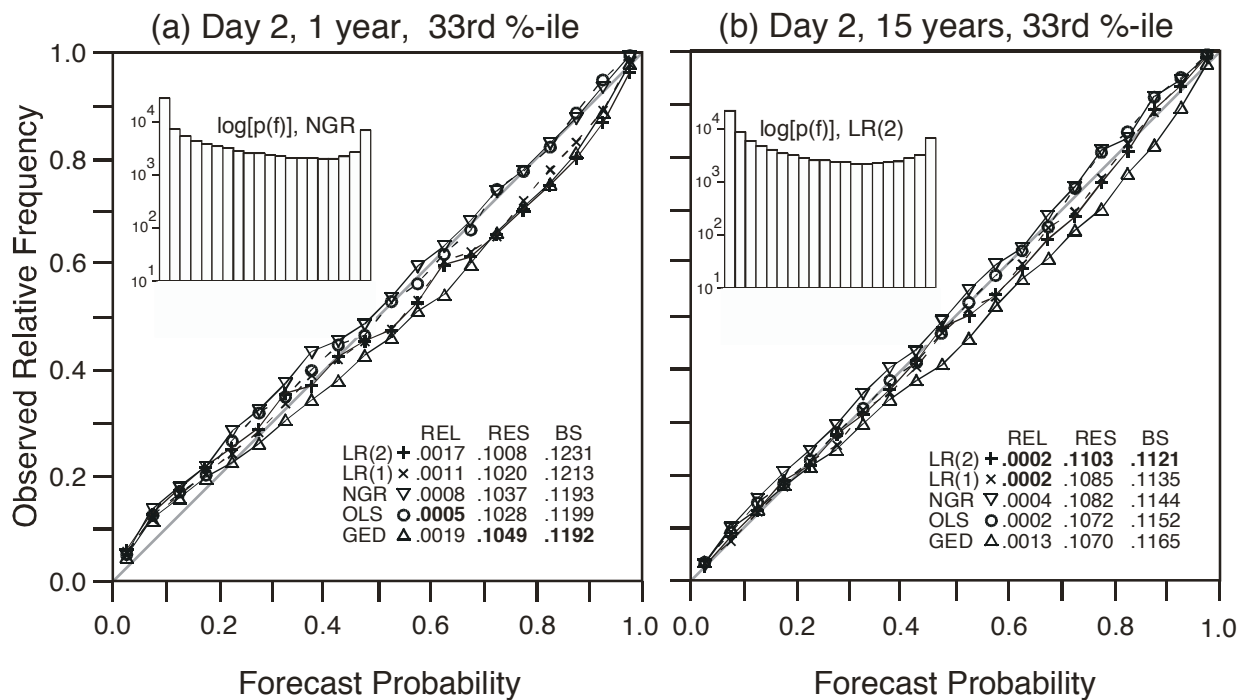


Figure 4. Reliability diagrams for day-2 October – March forecasts of the lower tercile of minimum temperature distributions, using (a) 1 year, and (b) 15 years of training data. Insets show frequencies of use of the forecasts for the best calibrated method in each case, and terms in the Murphy (1973) decomposition of the Brier score for each forecast method.

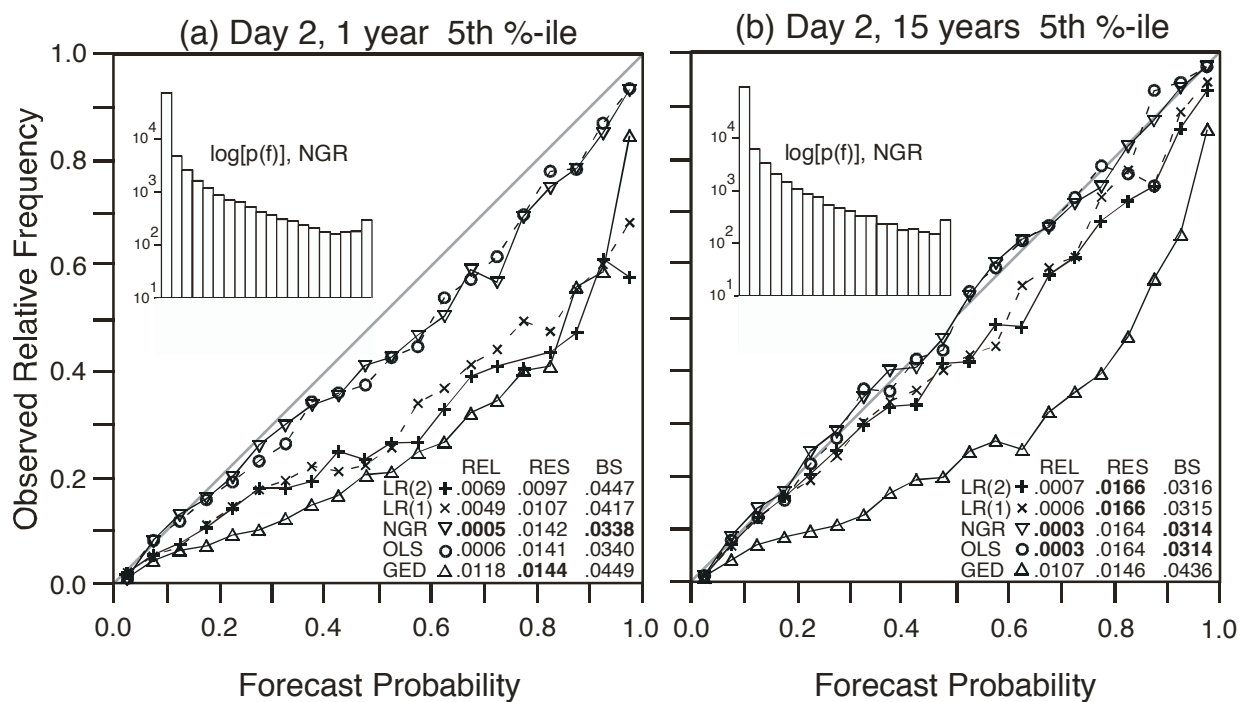


Figure 5. As Figure 4, for forecasts of the 5th percentile of cool-season minimum temperature.

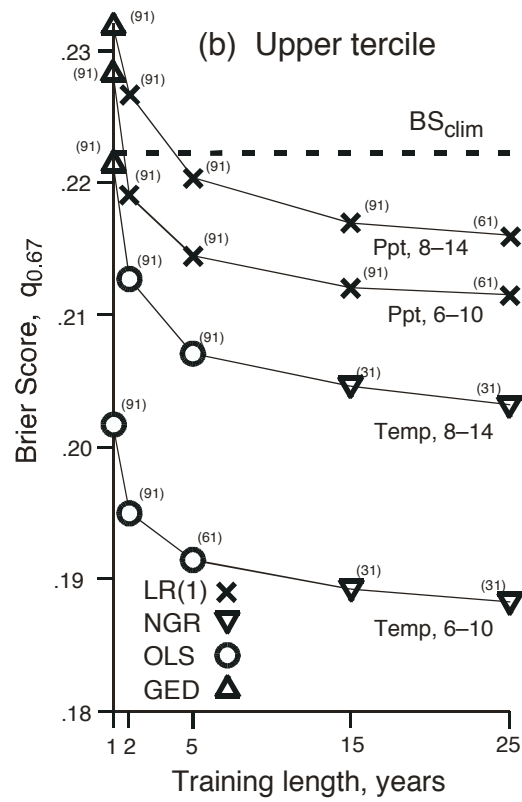
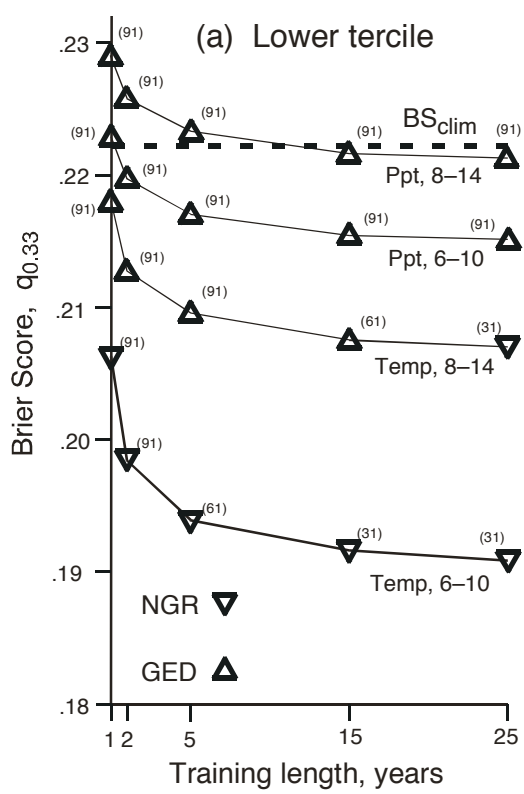


Figure 6. Brier scores for medium-range forecasts of outcomes below (a) the lower tercile, and (b) the upper tercile, of the climatological distributions, as functions of the length of the training data. Best forecast methods and training windows are shown in each case.

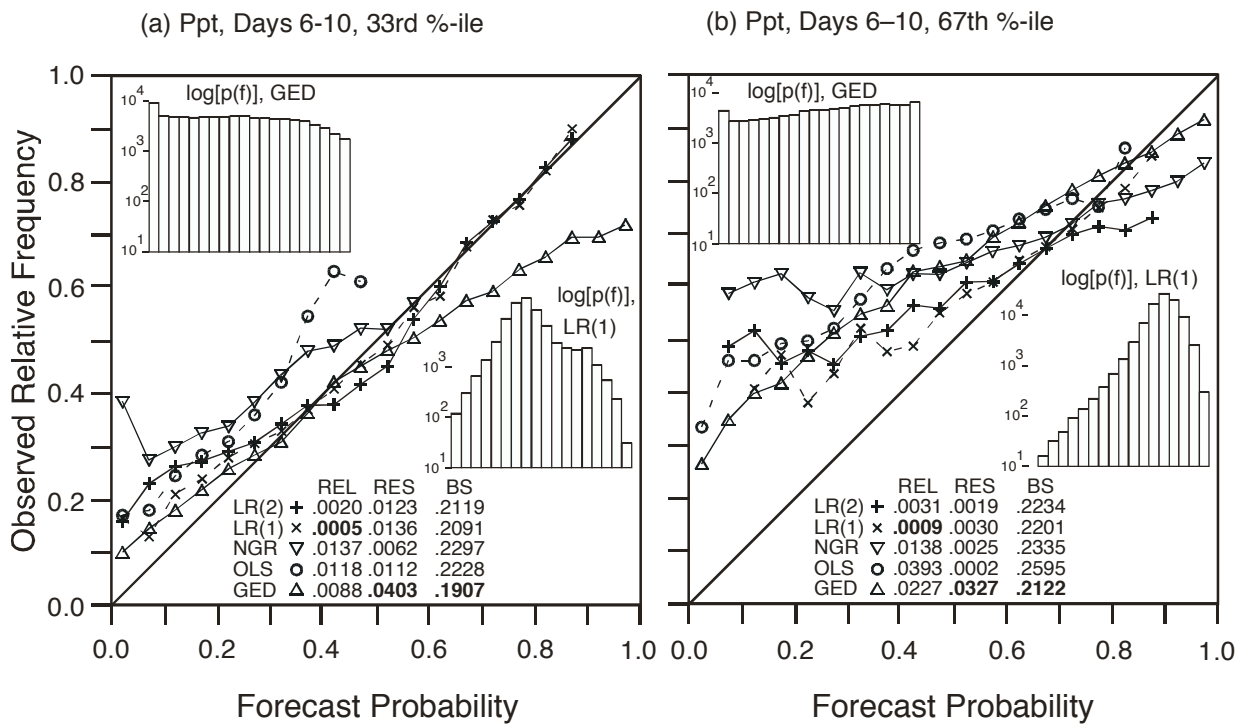


Figure 7. As Figure 4, for 6-10 day forecasts that precipitation is at or below the (a) lower terciles, and (b) upper terciles, using 15 years of training data.